Constructing disease onset signatures using multi-dimensional network-structured biomarkers

XIANG LI

Statistics and Decision Sciences, Janssen Research & Development, LLC, Raritan, NJ 08869, USA

DONGLIN ZENG

Department of Psychiatric, University of North Carolina, Chapel Hill, NC 27599, USA

KAREN MARDER, YUANJIA WANG*

Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 W 168th Street, New York, NY 10032, USA

yw2016@cumc.columbia.edu

SUMMARY

Potential disease-modifying therapies for neurodegenerative disorders need to be introduced prior to the symptomatic stage in order to be effective. However, current diagnosis of neurological disorders mostly rely on measurements of clinical symptoms and thus only identify symptomatic subjects in their late disease course. Thus, it is of interest to select and integrate biomarkers that may reflect early disease-related pathological changes for earlier diagnosis and recruiting pre-sypmtomatic subjects in a prevention clinical trial. Two sources of biological information are relevant to the construction of biomarker signatures for time to disease onset that is subject to right censoring. First, biomarkers' effects on disease onset may vary with a subject's baseline disease stage indicated by a particular marker. Second, biomarkers may be connected through networks, and their effects on disease may be informed by this network structure. To leverage these information, we propose a varying-coefficient hazards model to induce double smoothness over the dimension of the disease stage and over the space of network-structured biomarkers. The distinctive feature of the model is a non-parametric effect that captures non-linear change according to the disease stage and similarity among the effects of linked biomarkers. For estimation and feature selection, we use kernel smoothing of a regularized local partial likelihood and derive an efficient algorithm. Numeric simulations demonstrate significant improvements over existing methods in performance and computational efficiency. Finally, the methods are applied to our motivating study, a recently completed study of Huntington's disease (HD), where structural brain imaging measures are used to inform age-at-onset of HD and assist clinical trial design. The analysis offers new insights on the structural network signatures for premanifest HD subjects.

Keywords: Clinical trial design; Large-scale biomarkers; Locally varying effect; Network regularization; Neurological disorders.

*To whom correspondence should be addressed.

1. Introduction

It has been recognized in the neurological disorders research community that potential disease-modifying therapies will need to be introduced prior to the symptomatic stage. However, current diagnosis of neurological disorders rely on measurements of clinical symptoms and thus only identify symptomatic subjects. To assist in earlier diagnosis and ultimately early intervention, recent efforts have focused on revising current diagnostic criteria to incorporate objective biomarkers and subtle clinical signs that appear early in the disease process (Biglan *and others*, 2013). Early diagnosis may have acceptable sensitivity but low specificity when measured against current "gold standards" for clinical diagnosis. Thus, it is important to integrate the complementary contribution of biomarkers and estimate their effect profiles on time to disease diagnosis in order to improve accuracy. Furthermore, large-scale biomarkers (e.g., genomic and neuroimaging measures) are increasingly explored to assist clinical trial recruitment (Hua *and others*, 2016), which also calls for constructing biomarker signatures informative of a clinical trial endpoint (e.g., time-to-diagnosis).

Although Cox proportional hazards regression model has been used widely for time-to-event outcomes, it does not capture several real-world complexities. First, for many neurodegenerative disorders, biomarker effects may vary with age or diseases stage. For example, for Huntington's disease (HD), the CAG-repeat-Age Product score (CAP score, Zhang *and others*, 2011) is commonly used as an index summary marker for disease stage. As shown in Paulsen *and others* (2014) using data collected from our motivating study, PREDICT-HD, the changes in regional brain volumetric biomarkers, which are strongly associated with HD diagnosis, manifest different rates of decline for subjects at distinct disease stages represented by their CAP scores. The change in various markers as a function of CAP score is often non-linear (e.g., Figure 4 in Ross *and others*, 2014). See also our preliminary study in Section 2. As another example, research on neurobiological processes that underlie complex social and emotional behaviors reveals an age-dependent effect pattern for neuroimaging biomarkers. Hence, in order to accurately inform disease onset using these biomarkers, it is crucial to recognize and model their varying effects for subjects at a different disease stage or age.

Second, a large number of genomic and neuroimaging biomarkers are often collected, and biological studies may reveal meaningful network structures of the underlying relationship among the biomarkers. The motivating example for this work is the structural covariation network of brain (Chen *and others*, 2008; Eidelberg *and others*, 2011), which sheds lights on the patterns of structural changes in human brain. For such networks, the mean cortical thickness or other structural volumetric measures over each region of interest (ROI) are taken as the nodal value at the anatomical locations defined by the Automatic Anatomical Labeling atlas (Tzourio-Mazoyer *and others*, 2002) or the cortical atlas (Desikan *and others*, 2006). The network is estimated from association matrix of nodal values, and weighted edges are obtained from entries in the matrix, and unweighted and undirected edges are defined by thresholding the correlation matrix. Another example is the gene expression transcriptional network (Zhang and Horvath, 2005) characterized by co-expression of genes. When biomarkers are involved in the same biological process and thus affect diseases or phenotypes through some common biological mechanism or pathways, their effects manifested on disease outcome are expected to be correlated (Li and Li, 2010; Alexander-Bloch *and others*, 2013). Gene expressions and neuroimaging biomarkers often exhibit comparable effect trajectories due to their similar involvement in function (Zhang *and others*, 2013) or spatial proximity (Cuingnet *and others*, 2013).

This article aims to propose an efficient method to account for the aforementioned biologically relevant information when constructing biomarker signatures for time to disease onset. Incorporating multi-dimensional and network-structured biomarkers with local varying-effects poses complicated challenges on the analysis of time-to-event data. Existing methods can only handle some of the challenges but not all. To deal with large number of covariates for time-to-event outcomes, many penalized estimation methods

have been developed in recent years under linear model, non-linear model, and Cox model framework. See for example, Li and Liang (2008) for varying-coefficient models, Li and Li (2010) and Huang *and others* (2011) for linear regression, and Engler and Li (2009), Liu and Zeng (2013) for censored data. Other related work includes Zhang *and others* (2013) and Sun *and others* (2014) which assumes a constant vector of parameters and introduce Laplacian penalty to Cox model. None of the existing work can simultaneously accommodate moderate to large number of non-parametric varying-coefficient functions for predicting disease onset with network-structured biomarkers.

In this work, we adopt a local proportional hazards model framework for time-to-event outcomes to estimate biomarker profiles non-parametrically. There are two parallel smoothing procedures serving distinct goals. First, to reflect the underlying biology that subjects at a similar disease stage may share similar biomarker profile, biomarker effects are smoothed along the dimension of disease stage to achieve interpretable results matching the biological process. For this purpose, local kernel smoothing is used to borrow information across subjects with comparable disease stages. The second smoothing procedure assumes that the biomarkers linked in a network express correlated effects on the outcome, and hence their profiles are smoothed across the network links to borrow information from connected nodes. For this purpose and to incorporate large-scale biomarkers while achieving desirable sparseness, Laplacian penalty and L_1 penalty are implemented in the estimation. We develop a fast coordinate descent algorithm based on quadratic approximation at each local value of the disease stage. Additionally, we provide a data-driven approach to select tuning parameters with superior empirical performance.

The rest of this article is organized as follows. In the next section, we present the motivating study, PREDICT-HD (Paulsen *and others*, 2014), and some preliminary analyses. In Section 3, we describe a varying-coefficient network-regularized method under a local Cox proportional hazards model, a fast coordinate descent algorithm for implementation, and an effective procedure for the bandwidth selection. In Section 4, we show through simulation studies that our method significantly outperforms existing non-local approaches and demonstrate that our bandwidth selection procedure provides a tuning parameter close to the optimal value. In Section 5, we apply our methods to the motivating study, PREDICT-HD, where the whole brain structural magnetic resonance imaging (MRI) data are used to estimate a network-regularized biomarker signature for the age-at-onset of HD. The analysis results further reveal the differential effects of imaging biomarkers depending on a subject's baseline disease stage. Finally, we conclude the article with some discussions in Section 6.

2. Preliminary analyses of the motivating study data

There is an increasing body of literature suggesting that brain networks measured by neuroimaging biomarkers play important roles in the neurodegenerative process. For example, Chen *and others* (2008) demonstrated existence of topological patterns of cortical networks. Eidelberg *and others* (2011) described various kinds of brain networks in HD, and Feigin *and others* (2007) suggested thalamic metabolic spatial covariance is associated with age-at-onset of HD. Novak *and others* (2012) showed that structural connectivity-based topography of the basal ganglia is altered in premanifest and early manifest HD subjects. Furthermore, Alexander-Bloch *and others* (2013) suggested structural and functional connectivity are likely to be related, and thus the effects of connected nodes manifested on the clinical outcome may be similar. Incorporating such similarity over network structure in the analyses may offer greater accuracy to predict disease onset.

Another important empirical observation made in HD research (Zhang *and others*, 2011; Paulsen *and others*, 2014; Ross *and others*, 2014) is that the trajectories and effects of imaging and clinical biomarkers on the onset of HD may vary according to a widely used summary index of baseline disease stage/disease burden, referred as the CAP_s score (Zhang *and others*, 2011, details in Section 5). Thus, this CAP_s-dependent effect on HD onset can be captured by using CAP_s as the index of a varying-coefficient.

Our goal is to incorporate the baseline network structure and the prior evidence of local CAP-dependent effects to reflect the underlying biology and gain stability and efficiency for predicting age-at-onset of HD, especially in applications with large-scale data. We consider structural covariation network constructed from high-resolution structural MRI measures collected from healthy controls in the newly completed PREDICT-HD study (Paulsen *and others*, 2014). Baseline structural MRI measures were preprocessed (Paulsen *and others*, 2014) using Freesurfer 5.2 (http://surfer.nmr.mgh.harvard.edu), which provided delineation of cortical and subcortical ROIs. The covariation network structure is constructed based on the association (correlation) matrix of subcortical volumetric measures (Bullmore and Bassett, 2011).

To empirically illustrate the two main biological and clinical features of HD research (structural network and CAP_s-dependent effect), Figure A1(a) (Online supplementary material available at *Biostatistics* online) presents the structural covariation network estimated from the PREDICT-HD data. The structural covariation network shows that the left and right side ROIs of the same region are highly associated especially for Thalamus, Putamen, Amygdala, and Caudate regions. Using Thalamus ROI as an example, we fitted a non-parametric varying-effect hazard function model with age-at-onset of HD as outcome. Figure A1(b) (Online supplementary material available at *Biostatistics* online) shows that the effect profiles of these two highly linked ROIs ($\rho = 0.92$) are largely similar and vary with CAP_s.

In the remaining sections, we will propose methodology to construct network-informed (based on Figure A1(a) Online supplementary material available at *Biostatistics* online) local biomarker profiles to capture the changes according to CAP_s for all ROIs. Specifically, we will incorporate the profile similarity over structural covariation network as observed in Figure A1(a) (Online supplementary material available at *Biostatistics* online) and the varying effect along the dimension of CAP_s as observed in Figure A1(b) (Online supplementary material available at *Biostatistics* online). We expect that incorporating these two main biological features will improve the predictive performance for predicting HD onset in the presence of multi-dimensional neuroimaging biomarkers. We will also demonstrate the utilities of biomarker signatures to improve the clinical trial design efficiency. In addition, separating disease stage index CAP_s from other biomarkers in the network effectively accounts for their interactions. At the same time, the network structure among other biomarkers is preserved across all disease stages. To retain maximum flexibility, we do not assume any parametric form of the effect profiles. When a simpler functional form fits data adequately, our data-driven methodology will capitalize on this information to improve estimation efficiency.

3. Methodology for network-regularized varying coefficient hazards model

Let T_i be the time-to-event of interest (e.g., age-at-onset of a disease) and C_i be the censoring time for the ith subject. Denote by $\widetilde{T}_i = \min(T_i, C_i)$ the observed event time or censoring time and denote by $\delta_i = I(T_i \leq C_i)$ the event indicator, where $I(\cdot)$ is an indicator function. Let $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip})^T$ denote a vector of biomarkers and let W_i denote the index marker for the disease stage (e.g., CAP_s score representing the baseline disease stage for HD) which is treated separately from $\boldsymbol{X}_i$. We assume that given $\boldsymbol{X}_i$ and W_i, the censoring time is independent of event time, and observations collected on different subjects are independent. To capture the local effect from biomarkers that change with W, we consider a varying-coefficient Cox model, where the key interest is to estimate a vector of w-dependent functions (biomarker effect profiles as network-regularized signatures for disease onset), and a w-dependent baseline hazard function. Specifically, the hazard function takes the form

$$\lambda(t|\boldsymbol{X}_i, W_i) = \lambda_0(t, W_i) \exp\left(\boldsymbol{\beta}^T(W_i)\boldsymbol{X}_i\right),$$

where $\lambda_0(t, W_i)$ is an unspecified baseline hazard and $\boldsymbol{\beta}(w_0)$ is a vector of local effects at $W = w_0$.

Our methodology involves two parallel smoothing procedures. First, to borrow information from subjects with similar values of w to estimate $\boldsymbol{\beta}(\cdot)$ non-parametrically, we employ local kernel smoothing by incorporating a symmetric kernel function $K_{h_n}(\cdot)$ (Fan *and others*, 2006), where h_n is a data-dependent bandwidth. Consider the local partial likelihood at w_0

$$L(\boldsymbol{\beta}(w_0)) = \prod_{i=1}^{n} \left\{ \frac{\exp\left(\boldsymbol{\beta}^T(w_0)\boldsymbol{X}_i\right)}{\sum_{\widetilde{T}_k \geq \widetilde{T}_i} \exp\left(\boldsymbol{\beta}^T(w_0)\boldsymbol{X}_k\right) K_{h_n}(W_k - w_0)} \right\}^{\delta_i K_{h_n}(W_i - w_0)}$$

and notice the log-partial likelihood function of $\boldsymbol{\beta}(w_0)$ is given by

$$l(\boldsymbol{\beta}(w_0)) = \sum_{i=1}^{n} \delta_i K_{h_n}(W_i - w_0) \left\{ \boldsymbol{\beta}^T(w_0)\boldsymbol{X}_i - \log\left(\sum_{\widetilde{T}_k \geq \widetilde{T}_i} \exp\left(\boldsymbol{\beta}^T(w_0)\boldsymbol{X}_k\right) K_{h_n}(W_k - w_0) \right) \right\}. \tag{3.1}$$

Second, to incorporate network structure exhibited among components of $\boldsymbol{X}$ at each local point of w_0, the smoothness of $\boldsymbol{\beta}(w_0)$ over the network graph is induced by including a Laplacian penalty to the local likelihood. Specifically, since the biomarkers connected by an edge in the network express similar effects due to, for example, their similar involvement in structural or functional brain networks (Alexander-Bloch *and others*, 2013), or sharing genetic pathways in gene expression networks (Li and Li, 2010), we maximize the local log-partial likelihood with a Laplacian penalty term to induce smoothness of $\boldsymbol{\beta}(w_0)$ across network nodes. The estimator of $\boldsymbol{\beta}(w_0)$ is then obtained through

$$\widehat{\boldsymbol{\beta}}(w_0) = \underset{\boldsymbol{\beta}(w_0)}{\arg\min} \left\{ -\frac{1}{n} l(\boldsymbol{\beta}(w_0)) + p(\boldsymbol{\beta}(w_0); \lambda_1, \lambda_2) \right\} \tag{3.2}$$

and

$$\begin{aligned} p(\boldsymbol{\beta}(w_0); \lambda_1, \lambda_2) &= \lambda_1 \|\boldsymbol{\beta}(w_0)\|_1 + \frac{\lambda_2}{2} \boldsymbol{\beta}^T(w_0) \boldsymbol{L} \boldsymbol{\beta}(w_0) \\ &= \lambda_1 \sum_k |\beta_k(w_0)| + \frac{\lambda_2}{2} \sum_{(k,l) \in E} v_{kl} \left(\frac{\beta_k(w_0)}{\sqrt{d_k}} - \frac{\beta_l(w_0)}{\sqrt{d_l}} \right)^2, \end{aligned} \tag{3.3}$$

where $\|\cdot\|_1$ is the L_1 norm to induce sparsity, E is the set of all edges, w_{kl} is the given weight between nodes k and l ($v_{kk} = 0$), d_k is the degree of the k^{th} node, and the entry (k, l) of Laplacian matrix $\boldsymbol{L}$ is defined as

$$L_{kl} = \begin{cases} 1 - v_{kk}/d_k, & \text{if } k = l \text{ and } d_k \neq 0, \\ -v_{kl}/\sqrt{d_k d_l}, & \text{if } k \text{ and } l \text{ are adjacent,} \\ 0, & \text{otherwise.} \end{cases}$$

Node k is an isolated node if $d_k = 0$.

Remarks. The weights v_{kl} usually represent the strength of the link between nodes k and l. Conventionally, anatomical connectivity, which is inferred from thresholding an association matrix, yields $v_{kl} = 1$ if the association between them is larger than a given threshold and 0 otherwise (Bullmore and Bassett, 2011). In some situations, continuous weights v_{kl} may be used and $d_k = \sum_{l:(k,l) \in E} v_{kl}$ is adopted as a way to measure the degree of a node without applying any threshold to the association matrix. In our motivating example in Section 2, the network is defined by the association matrix of structural MRI measures and the

continuous weights are used. Lastly, we remark that local Lasso and Enet penalties are two special cases of (3.3): when $\lambda_2 = 0$ or $\boldsymbol{L}$ is an identity matrix, respectively.

To obtain the profile of coefficients, we solve the objective function (3.2) at a series of w_0 for $w_0 \in W$, where W could be a set of all data points, equally spaced points over a range or any user-defined points:

$$\widehat{\boldsymbol{\beta}}(W) = \underset{\boldsymbol{\beta}(w_0):w_0\in W}{\arg\min} \left\{ \sum_{w_0\in W} \left(-\frac{1}{n} l(\boldsymbol{\beta}(w_0)) + p(\boldsymbol{\beta}(w_0); \lambda_1, \lambda_2) \right) \right\}. \tag{3.4}$$

Each $\boldsymbol{\beta}(w_0)$ is estimated separately but with the same set of tuning parameters. Additionally, we propose an adaptive version of the Laplacian network penalty (AlocNet). In this version, we modify the L_1 penalty to an adaptive version as in Zhang and Lu (2007) and also adjust the signs of the coefficients in the Laplacian penalty to better handle the linked covariate with opposite signs of the effect sizes. Similar approach of accounting for the opposite signs of parameters in a network was proposed in Kunegis *and others* (2010), Zhang *and others* (2012), and Sun *and others* (2014). In our case, the local penalty term at w_0 in AlocNet is

$$\begin{aligned} &p^*(\boldsymbol{\beta}(w_0); \breve{\boldsymbol{\beta}}(w_0), \lambda_1, \lambda_2) \\ &= \lambda_1 \sum_i \frac{|\beta_i(w_0)|}{|\breve{\beta}_i(w_0)|} + \frac{\lambda_2}{2} \sum_{(k,l)\in E} w_{kl} \left(\frac{\text{sign}(\breve{\beta}_k(w_0))\beta_k(w_0)}{\sqrt{d_k}} - \frac{\text{sign}(\breve{\beta}_l(w_0))\beta_l(w_0)}{\sqrt{d_l}} \right)^2, \end{aligned}$$

where $\breve{\boldsymbol{\beta}}(w_0)$ is an initial estimator of $\boldsymbol{\beta}$. It could be obtained from a ridge regression by replacing $p(\boldsymbol{\beta}(w_0); \lambda_1, \lambda_2)$ in (3.2) with $\lambda \sum_{i=1}^p \beta_i^2(w_0)$.

Key advantages of the proposed method are the induced double smoothness over the dimension of disease stage (e.g., W) and the nodes in the network space, the ability to incorporate their potential non-linear interactions, and the capability to handle large number of structured biomarkers while producing a sparse and interpretable model. With the estimated $\boldsymbol{\beta}(w_0)$, we can estimate the baseline cumulative hazard function at w_0 as

$$\widehat{\Lambda}(t, w_0) = \sum_{i=1}^n \frac{\delta_i I(\widetilde{T}_i \le t) K_{h_n}(W_i - w_0)}{\sum_{\widetilde{T}_k \ge \widetilde{T}_i} \exp\{\widehat{\boldsymbol{\beta}}(w_0)^T X_k\} K_{h_n}(W_k - w_0)}. \tag{3.5}$$

We present a fast coordinate descent algorithm based on quadratic approximation at each value of the disease stage for computation in Section A1 of the Supplementary material available at *Biostatistics* online. The details on tuning parameter selection and computational efficiency analysis are described in Sections A2 and A3 of the Supplementary material available at *Biostatistics* online. We prove the theoretical oracle property of the proposed method in Section A4 of the Supplementary material available at *Biostatistics* online. The main challenge is that the non-parametric approximation bias (from local kernel smoothing) needs to be controlled carefully at certain stochastic rates, which entails using non-parametric regression theories and large deviation results to derive weak oracle properties for a class of network-regularized optimization problems.

4. Simulation studies

We conducted extensive simulations to evaluate the performance of the proposed local varying-coefficient network method with various penalty functions for variable selection. To mimic the structural covariation network structure, we constructed $\boldsymbol{X}$ in independent blocks/regions and the nodes within each block

were correlated with each other. We simulated the network to be consisted of 23 blocks and each block contained five nodes: for the kth block of $\boldsymbol{X}$ ($k = 1, ..., 23$), we generated $\boldsymbol{X}_k$ from a multivariate normal distribution $\boldsymbol{X}_k \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{P})$, where $\boldsymbol{P}$ was a correlation matrix with ρ as off-diagonal entries. Three degrees of correlations, $\rho = 0.2, 0.5, 0.8$, were considered to evaluate different strengths of network structures. Only 15 covariates from three blocks had non-zero effects on the outcome, and the effects of the informative biomarkers varied according to the disease stage marker, W (e.g., CAP score for predicting HD onset), which was generated from a uniform distribution $U(0, 1)$. The underlying hazard function for subject i was given by $\lambda(t|\boldsymbol{X}_i, W_i) = \lambda_0(t) \exp\left\{\sum_{j=1}^{15} \beta_j(W_i)X_{ij}\right\}$, where $\lambda_0(t)$ was specified by a Weibull distribution with shape parameter 5 and scale parameter 2. Censoring time was generated from the uniform distribution to yield 30% censored subjects. For $\beta_j(w)$'s, two different models were considered in the simulation. For the first model, we specified the coefficients as $\beta_1(w) = \cdots = \beta_5(w) = 2\exp(-w^3), \beta_6(w) = \cdots = \beta_{10}(w) = 4w(1-w), \beta_{11}(w) = \cdots = \beta_{15}(w) = \sin(\pi w) + \cos(\pi w)$. In the second model, we considered the case when $\boldsymbol{\beta}(w)$ had the same forms as in the first model but within each block, two of the five nodes were assigned opposite signs.

We considered the sample size $n = 500$. For each simulated dataset, we considered the local points $W_0 = \{w_0\}$ with w_0 from 0.1 to 0.9 with increments of 0.1, the grid of α between 0.1 to 0.9 with increments of 0.1, the length of searching path for λ at $K = 20$ and the bandwidth at $h = 0.1, \cdots, 0.5$. Ten-fold cross-validation was applied to choose the optimal tuning parameters and the bandwidth based on cross-validated partial likelihood. For each model, simulations were repeated 100 times.

In all simulation studies, we reported the estimation accuracy, variable selection performance, and the data-driven procedure to select the bandwidth. We computed integrated mean squared error (IMSE) to evaluate the estimation performance with

$$\text{IMSE} = \sum_{j=1}^{15} \sum_{w_i \in W_0} \frac{(\hat{\beta}_j(w_i) - \beta_j(w_i))^2}{|W_0|}.$$

We calculated the accuracy (ACC) defined as the percentage of correctly identified relevant and irrelevant covariates, the number of incorrectly identified non-zero covariates (IN), and the number of correctly identified non-zero covariates (CN).

Table 1 summarizes simulation results comparing the proposed local methods with existing methods under various penalty functions (e.g., Anet, Net, Enet, Lasso). When $\boldsymbol{\beta}(w)$ is assumed to be constant over the range of w (non-local methods) and using the adaptive network penalty (Anet), our method can be reduced to the non-local methods in Zhang *and others* (2013) and Sun *and others* (2014). Table 1 shows that the proposed localized methods (e..g., AlocNet) performed better than non-localized ones in terms of both estimation accuracy and variable selection property due to taking advantage of the local smoothing. Within the localized methods, we observed that AlocNet gave the smallest IMSE for both Models 1 and 2 with different correlations among the network covariates. AlocNet substantially reduced the IMSE compared to all alternatives (e.g., around 60% reduction compared to locNet). When the connected biomarkers have the same signs, locNet performed better than locEnet and locLasso. For Model 2 when the signs were different for the connected biomarkers, locNet had worse performance compared to AlocNet because the unadjusted penalty (3.3) cannot reflect the opposite signs. LocNet performed similarly as locEnet and better than locLasso when the correlation was medium and high.

For the variable selection performance, we calculated ACC, IN, and CN. All four local methods were able to correctly identify both informative and null covariates except for Model 2 with high correlation among the covariates. For Model 1, AlocNet, and locNet had similar performance and better than locEnet and locLasso with slightly more correctly identified non-zero and less incorrectly identified non-zero covariates. For Model 2, AlocNet had the best performance on variable selection. When the correlation

Table 1. *Comparison of estimation and selection performance using proposed local methods (i.e., local adaptive network [AlocNet], local network [locNet], local elastic-net [locEnet], and local lasso [locLasso]) and existing non-local methods (adaptive network [ANet], network [Net], elastic-net [Enet], and Lasso) for Models 1 and 2 with three different network structures* $\rho = 0.2, 0.5, 0.8$

	Proposed local methods				Non-local methods			
	AlocNet	locNet	locEnet	locLasso	ANet	Net	Enet	Lasso
				Model 1				
				$\rho = 0.2$				
IMSE*	2.32	5.42	8.60	8.48	13.40	14.36	14.64	14.52
ACC†	1.00	1.00	1.00	1.00	0.89	0.80	0.82	0.85
IN‡	0.00	0.00	0.02	0.02	11.38	23.35	20.96	16.74
CN§	15.00	15.00	14.94	14.95	14.27	14.98	14.76	14.72
				$\rho = 0.5$				
IMSE	2.18	6.16	7.97	8.51	15.11	15.69	15.92	15.85
ACC	1.00	1.00	1.00	1.00	0.89	0.86	0.85	0.91
IN	0.00	0.00	0.00	0.00	10.57	15.88	16.18	9.44
CN	15.00	15.00	14.93	14.91	12.74	14.78	14.33	14.01
				$\rho = 0.8$				
IMSE	2.35	5.00	8.40	8.38	16.27	16.46	16.62	16.64
ACC	1.00	1.00	1.00	1.00	0.87	0.89	0.87	0.92
IN	0.00	0.00	0.03	0.07	10.38	12.55	13.82	6.21
CN	14.99	15.00	14.90	14.78	10.28	14.38	13.72	12.39
				Model 2				
				$\rho = 0.2$				
IMSE	2.52	8.22	8.16	7.80	11.06	12.82	12.82	12.70
ACC	1.00	1.00	1.00	1.00	0.89	0.76	0.76	0.77
IN	0.00	0.06	0.04	0.08	12.10	27.36	27.58	25.84
CN	15.00	14.86	14.86	14.87	14.86	14.92	14.92	14.89
				$\rho = 0.5$				
IMSE	2.28	7.93	7.47	8.22	10.03	12.01	11.99	11.88
ACC	1.00	0.99	1.00	0.99	0.88	0.75	0.74	0.76
IN	0.00	0.20	0.22	0.26	13.99	28.75	29.17	27.51
CN	14.92	14.38	14.67	13.88	14.64	14.80	14.82	14.74
				$\rho = 0.8$				
IMSE	3.17	9.94	9.09	11.29	8.79	11.12	11.11	11.08
ACC	0.99	0.97	0.97	0.96	0.83	0.72	0.72	0.74
IN	0.40	0.36	0.24	0.73	18.87	31.06	31.14	28.61
CN	14.26	11.61	11.96	11.31	14.10	13.96	13.97	13.82

*Integrated mean square error.

†Percentage of correctly identified relevant and irrelevant covariates.

‡Number of incorrectly identified non-zero covariates.

§Number of correctly identified non-zero covariates.

is high, AlocNet outperforms all other alternatives in terms of corrected identified informative covariates. Note that non-local methods have reasonable performance on CN, but suffers from producing non-sparse model.

Table 2. *Performance of the bandwidth selection procedure for the proposed local methods (i.e., local adaptive network [AlocNet], local network [locNet], local elastic-net [locEnet], and local lasso [locLasso])*

	Model 1				Model 2			
	AlocNet	locNet	locEnet	locLasso	AlocNet	locNet	locEnet	locLasso
		$\rho = 0.2$				$\rho = 0.2$		
Best*	0.12	0.12	0.23	0.21	0.21	0.27	0.27	0.25
Selected	0.12	0.18	0.24	0.22	0.17	0.32	0.31	0.32
		$\rho = 0.5$				$\rho = 0.5$		
Best	0.10	0.10	0.19	0.19	0.21	0.30	0.30	0.30
Selected	0.10	0.12	0.22	0.21	0.19	0.36	0.37	0.35
		$\rho = 0.8$				$\rho = 0.8$		
Best	0.10	0.11	0.14	0.12	0.26	0.34	0.34	0.34
Selected	0.10	0.11	0.19	0.19	0.22	0.34	0.35	0.30

*Defined as the one with smallest IMSE.

Next, we investigate performance of the proposed bandwidth selection procedure. For each replication, we obtained the estimates for $h = 0.1, \ldots, 0.5$, and defined the "best" bandwidth as the one with the smallest IMSE. In Table 2, we observe that the selected bandwidths based on our procedure were very close to the "best" bandwidths, indicating satisfactory performance of our data-driven procedure.

To make the method easily accessible for routine data analysis, we have developed an R-package "Coxnet" and implemented several techniques for efficient computation: (i) use warm starts for estimating $\boldsymbol{\beta}(w_0)$ along a regularization path; (ii) use one-step coordinate descent to save time in the intermediate update steps; and (iii) use sparse data structure to save computer memory and the time to search for the non-zero coefficients in a sparse $\boldsymbol{\beta}(w_0)$. A running time analysis is performed for simulation setting Model 1 with various sample sizes and numbers of parameters. For each case, we set the entire solution path with 20 values and compute 9 data points in W with 10-fold cross-validation. The running time for completing the entire solution path for all values of W and with cross-validation are summarized in the Section A3 of the Supplementary material available at *Biostatistics* online. From Figure A1 of the Supplementary material available at *Biostatistics* online, we can see that the running time increases relatively linear with increasing n and p.

5. Application to predict study

HD is a severe neurodegenerative disorder caused by expansion of trinucleotide "C-A-G" repeats at the *HTT* gene (MacDonald *and others*, 1993). All subjects with expanded repeats will eventually be diagnosed with HD, but when will disease onset occur in lifetime remains unknown. We focus on the identification of biomarkers preceding clinical diagnosis and informative of age at diagnosis to understand disease mechanism and assist the design of neuroprotective clinical trials.

As observed in Figure **??** and suggested from the literature, the effects of HD biomarkers vary according to disease stage or disease burden defined by the CAP score (Zhang *and others*, 2011). CAP score is constructed from a centered product of "C-A-G" repeats length and age. Standardized CAP is referred as CAP_s. A *Low-Med-High* risk of onset classification was suggested in the literature (Zhang *and others*, 2011): $\text{CAP}_s = 0.67$ (low CAP_s, corresponds to unstandardized CAP $= 290$) and 0.85 (medium CAP_s,

corresponds to unstandardized CAP = 367) for distinguishing subjects at low risk and medium risk of onset, and $CAP_s = 1.00$ (corresponds to CAP = 432) is another critical value, indicating high risk. We also assessed the proportional hazard assumption by grouping subjects based on the quartiles of CAP score W and tested the assumption in each group. The results show non-significance for all groups (P-value = 0.67, 0.86, 0.86, 0.74 for each group), which may imply the assumption is valid in our application.

We applied eight methods described in Section 4 to 840 subjects at risk of HD, who participated in the PREDICT study and did not have HD at baseline exam. The median follow-up time was 3 years (up to 8 years), and 128 subjects developed HD during the study. Variable W of the varying coefficient is the baseline CAP_s. The feature variables $\boldsymbol{X}$ include 8 non-imaging demographic and clinical variables [e.g., gender, education, baseline total motor score (TMS) from the Unified Huntington's Disease Rating Scale (UHDRS), and cognitive and functioning measures] and 28 subcortical MRI imaging ROI biomarkers measured at the baseline visit. Ten-fold cross-validation was used to choose the tuning parameters and bandwidth. All variables were standardized before model fitting.

Among the non-local methods, ANet, Net, Enet, and Lasso selected 19, 31, 30, and 24 biomarkers, respectively. ANet had the highest C-index (Harrell *and others*, 1982) of 0.83 compared to Net, Enet, and Lasso, indicating better performance, and we report some of its estimated coefficients in Table 3. As a comparison, a regular Cox model without variable selection including all markers was fit and found to yield a C-index of 0.84, but failed to identify several important imaging markers reported in other literature such as Caudate and Putamen (Paulsen *and others*, 2014), potentially due to its inability to handle correlated imaging biomarkers.

To be succinct, for the local methods we report the numeric estimates from AlocNet, locNet, locEnet, locLasso at three values defining low-, medium-, and high-disease-burden group ($CAP_s = 0.67, 0.85, 1.00$, respectively, Section 2). Table 3 presents the biomarkers identified by AlocNet. Comparing non-local method ANet with AlocNet, we can see that ANet misses some important biomarkers previously reported in the literature (Paulsen *and others*, 2014), such as Symbol Digit Modality Test (SDMT), and does not detect the varying effects with CAP_s. The C-index for AlocNet is much higher than Anet at all three risk groups. Most image biomarkers express greater effects for medium- and high-risk subjects as estimated the local methods. These results suggest that imaging markers' effects depend on the baseline disease burden, justifying using a varying coefficient network-regularized model with CAP_s as an index biomarker W.

In terms of feature selection and prediction performance, among the four local methods, Table 3 shows that LocLasso selects the least number of biomarkers and has worse prediction performance. LocNet and LocEnet give similar estimation and prediction performance, which is consistent with the simulation results in the presence of opposite signs in Table 1. Compared to locNet and locEnet which do not use the network Laplacian penalty $\boldsymbol{L}$, AlocNet has comparable prediction performance at low CAP_s but better performance at medium and high CAP_s with only slightly more biomarkers included. This comparison illustrates the advantage of incorporating structural covariation network in prediction by AlocNet.

It may be of interest to note that some of the left and right side of brain presents asymmetric effects. For example, only one side is identified for most of imaging biomarkers except Pallidum and Thalamus. In addition, only right side of Pallidum shows an effect at low CAP_s while the left side has effects at medium and high score, suggesting that the right side Pallidum degeneration might be more predictive for the early stage and the left side may be more predictive at medium and high score. Asymmetric neurodegeneration in HD in striatal regions reported previously in Rosas *and others* (2001) was the right side first. However, HD has been largely considered as a symmetric disease, the biological implication of asymmetry is unclear, and these findings need to be confirmed in future studies.

Figure 1 graphically displays the profiles of the top 10 clinical variables and imaging biomarkers identified by AlocNet as a function of baseline CAP_s, and their 95% confidence intervals based on bootstrap method. The two vertical dashed lines represent the thresholds for the *Low-Med-High* classification in (Zhang *and others*, 2011) based on levels of CAP_s. The "Clinical markers" subfigure in Figure 1 shows

Table 3. *Estimated biomarker effects from proposed local methods (i.e., local adaptive network [AlocNet], local network [locNet], local elastic-net [locEnet], and local lasso [locLasso]) at* CAP_s = *Low* (0.67), *Med* (0.85), *High* (1.00) *and existing non-local method (adaptive network [ANet])*

	AlocNet			locNet			locEnet			locLasso			ANet
CAP_s	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High	
TMS*	0.49	0.67	0.45	0.40	0.63	0.41	0.40	0.63	0.41	0.45	0.67	0.40	0.41
SDMT*	−0.35												
Thalamus, L*	0.19	0.48	0.30	0.10	0.30	0.39	0.10	0.31	0.40		0.32	0.48	0.35
Caudate, L*		−0.39	−0.14										
Pallidum, L*	−0.50	−0.61	−0.08	−0.30	−0.43	−0.08	−0.31	−0.43			−0.31	−0.49	−0.68
Amygdala, L*	0.11	0.28	0.07										0.19
Cerebellum, R*	−0.09	−0.33											−0.13
Thalamus, R*	0.16	0.51	0.55	0.13	0.35	0.34	0.13	0.36	0.34		0.31	0.28	0.42
Putamen, R*		−0.30	−0.42										−0.13
Pallidum, R*	−0.42	−0.26	−0.14	−0.05	−0.27	−0.14	−0.05						−0.01
Accumbens, R*		0.33	0.20										
Non-zero†	7	9	8	6	6	6	6	6	6	1	5	5	19
C-index†	0.96	0.90	0.95	0.96	0.86	0.91	0.96	0.86	0.90	0.69	0.88	0.89	0.83

L, left side; R, right side.

*Identified by AlocNet.

†based on all selected markers in each model.

that both TMS and SDMT affect HD onset for early stage subjects (low CAP_s) while the effect from TMS persists for the medium- and high-risk stage subjects (medium and high CAP_s) but the effect from SDMT disappears. The rest "Imaging biomarkers" subfigures show that most subcortical imaging biomarkers manifest stronger effects as the disease burden, CAP_s, increases. For low risk subjects, for example with CAP_s=0.67, only Pallidum-R (volume of the right side of Palladium region) has an effect significantly different from zero. For medium- and high-risk subjects, more biomarkers become active. These results further demonstrate differential effects of imaging markers depending on CAP_s, thus the differential utilities to serve as risk markers to predict future disease onset. For example, Pallidum-R is useful for early- and medium-stage subjects, and Putamen-R, Pallidum-L, Amygdala-L, Accumbens-R, Caudate-L, Cerebellum-R, and Thalamus are useful for high risk subjects. Paulsen *and others* (2014) documented various measures associated with HD progression representing imaging, motor, cognitive, functional, and psychiatric domains. They showed different rates of decline between premanifest HD and controls using longitudinal data and provided a ranking for several clinical and imaging biomarkers. In this analysis, we identified similar sets of important imaging biomarkers, but with much more detailed characterization of their effect profiles over baseline disease stage.

In Figure 2, we present the imaging networks and their effects estimated by AlocNet at $CAP_s = 0.67, 0.85, 1.00$. Since the covariation network structure was estimated from the same underlying control population, the estimated profiles represent the effects of ROIs on HD onset at various baseline disease stage. For graphical demonstration purpose, a node and its edges were removed from the figure if the node had no effect on disease onset as estimated by AlocNet. We observe that the network effects change with CAP_s. For low risk subjects with $CAP_s = 0.67$, a highly connected network ($\rho \geq 0.7$) for HD onset is identified between Thalamus and Pallidum, similar to Feigin *and others* (2007). The topology of one network signature including Thalamus, Caudate, Pallidum, and Putamen remains the same for medium and high risk subjects, but with different effect size (radius of nodes). As CAP_s increases from 0.85 to 1.00, the effect sizes of both Thalamus-R and Putamen-R increase which may be due to the strong connection between them, while the effect sizes of Caudate-L and Thalamus-L decreases. Note that Thalamus and Putamen regions were also identified as structural covariance patterns by Feigin *and others* (2007) in a principal component analysis using positron emission tomography imaging measures.

Lastly, we illustrate the clinical significance of our approach through using the AlocNet-identified biomarker signatures to design a hypothetical neuroprotective prevention clinical trial with time-to-HD diagnosis as the endpoint. We compare a targeted recruitment scheme that enriches the sample by including subjects with a positive biomarker signature at the baseline (e.g., risk score $\boldsymbol{\beta}^T(w)\boldsymbol{X}$ greater than the median of the population risk scores). The probability of receiving a HD diagnosis within next 5 years for a premanifest subject with *HTT* gene mutation (expanded "C-A-G" repeats, Section 2) was estimated using $\widehat{\boldsymbol{\beta}}(w)$ and $\widehat{\Lambda}(t, w)$ fitted by AlocNet. Since PREDICT-HD was a natural history study where no subject had received an active intervention, these estimates provide design parameters for the placebo arm in a clinical trial. Figure S.2 of the Supplementary material available at *Biostatistics* online shows the required sample size as a function of a range of given effect sizes (assumed hazard ratio of the intervention). The targeted recruitment strategy substantially reduces required sample size, for example, from more than 17 000 to close to 3000 in low risk group assuming a hazard ratio of 0.8 of the intervention (i.e., a 20% reduction in hazard in the intervention group compared to the placebo group), and from about 5000 to less than 3000 for high risk subjects. The improvement is greater for smaller effect sizes and low risk group.

In summary, differential smooth effects of imaging ROI networks depending on CAP_s are observed in our analyses. The network signature for low risk subjects involves regions for core motor control information hub (e.g., Thalamus, Pallidum), while additional regions for fine motor control and coordination (e.g., Putamen) and reward system (e.g., Amygdala, Accumbens) become active for medium and high risk subjects. The medium- and high-risk networks share similar topological structure but with different effects. Recruiting subjects with a positive biomarker signature substantially improves clinical trial efficiency.

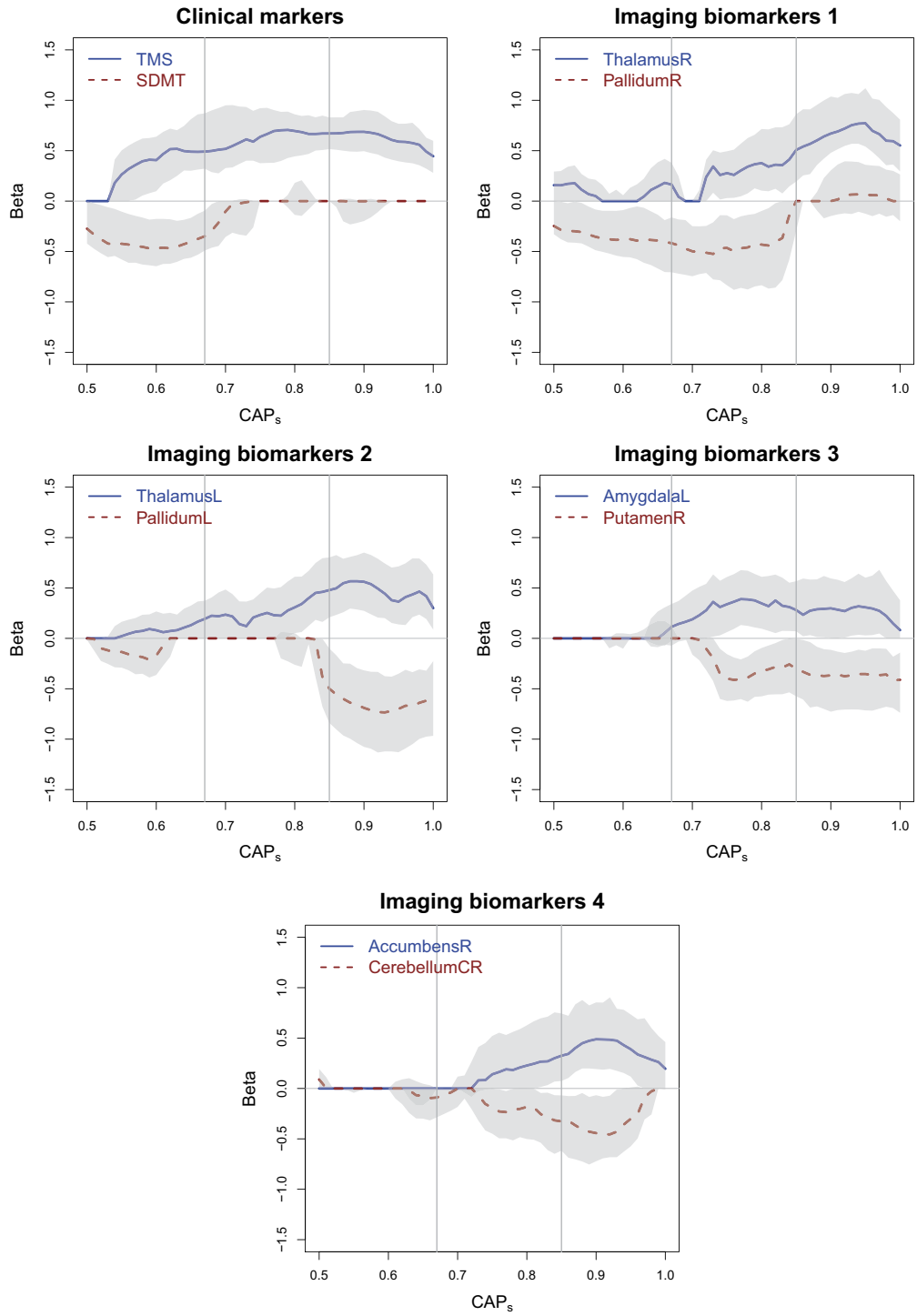

Fig. 1. Estimated effect profiles of the top 10 biomarkers (2 clinical markers and 8 imaging ROI markers) on the hazard of HD identified by the proposed local adaptive network [AlocNet] as a function of disease stage CAP_s. Shaded area indicates pointwise 95% confidence interval. Two vertical lines divide early, moderate, and late stage based on CAP_s. Few markers have an effect at the early stage, while more markers manifest an effect at the medium or late disease stage.

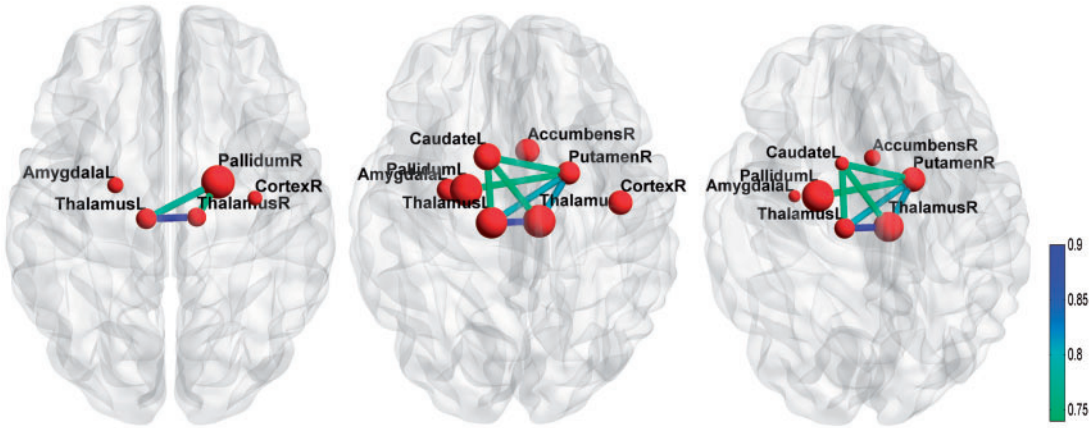

Fig. 2. Subcortical network signatures for age-at-onset of HD using PREDICT baseline imaging data estimated by local adaptive network [AlocNet]. Node radius indicating standardized effect of a node on HD onset.

6. Discussion

In this work, we propose a method to estimate locally varying biomarker network signature for disease age-at-onset. Our method can handle large-scale structured biomarkers without imposing a parametric form of varying-effect profiles, and thus cover both linear and potentially non-linear interactions between network biomarkers and the disease stage biomarker. The procedure also effectively selects important biomarkers with opposite signs due to incorporating sign information in the Laplacian penalty matrix. The real data application reveals insights on CAP_s-dependent structural MRI network signatures for HD onset.

A simpler alternative to our method is to binning data by disease stage and then fitting a separate model for each bin, which would provide a crude estimate of the effect profiles $\boldsymbol{\beta}(w)$ at certain bins of w. In fact, this is equivalent to fitting a piece-wise constant function of effect profiles. However, such a method would bear large bias and does not produce smooth functions $\boldsymbol{\beta}(w)$. Using a higher order kernel as compared to lower order kernel (binning) will improve optimal rate if the true underlying effect function is smooth.

For linear regression, Huang *and others* (2011) showed that incorporating Laplacian penalty reduces variance without incurring any bias on the estimator. Since locally around the true parameters, our objective function is approximated by a quadratic function, similar results may hold for our method. In practice, when there is a lack of biological knowledge to suggest effectiveness of a Laplacian prior, it can be checked by fitting a model without such penalty and examining the resulting biomarker profiles. Similar profiles for highly linked variables suggest that incorporating a Laplacian prior is sensible for improving prediction of outcomes and reducing overfitting with multi-dimensional biomarkers. As shown in Figure **??**, when fitting the model without Laplacian penalty using PREDICT data, the effects of left and right side of Thalamus ($\rho = 0.92$) are largely similar. Furthermore, the tuning parameter associated with the Laplacian prior is selected in a data-adaptive fashion; therefore, a large tuning parameter (smaller α) value will indicate the usefulness of the Laplacian prior for accounting the network. Particularly, in our data application, this tuning parameter is $\alpha = 0.1$.

In our application, structural covariation network was used in the Laplacian penalty. In other applications, networks based on partial correlation that adjusts for confounding by other biomarkers can be used when deemed appropriate. Another extension to consider is to obtain a localized network structure. Subjects in different disease stages at the baseline exam (e.g., with different CAP scores, w_0) may exhibit

distinct network patterns. It is straightforward to incorporate local network structure in our procedure by re-defining penalty function to be dependent on the disease stage CAP score w_0: $\boldsymbol{\beta}^T(w_0)\boldsymbol{L}(w_0)\boldsymbol{\beta}(w_0)$, where $\boldsymbol{L}(w_0)$ is the local Laplacian matrix. Incorporating spatial correlation to estimate spatial–temporal covariance network is also of interest, and some latent graph models may be explored.

Here, we did not study causal network such as gene regulatory network (Kanehisa and Goto, 2000), where causal pathways and directions between genes may be known. In these cases, using a directed acyclic graph to characterize network may be more appropriate than a Laplacian prior. However, in the case of HD, since the *HTT* gene is dominant, while there may be other genes that modify age-at-onset, they are very rare and effects may be modest (Lee *and others*, 2015). We assumed the network structure is available from external sources or healthy control population. A challenging next step would be to estimate network structure from data and incorporate estimation uncertainty in the inference for the biomarker profiles. Lastly, when longitudinal measurements on biomarkers are available, change of network may be considered to predict clinical outcomes.

7. Software and data

We provide an efficient R implementation of our algorithms in the R-package "Coxnet" available at GitHub (https://github.com/yuanjiawang/Coxnet) and CRAN (https://cran.r-project.org/web/packages/Coxnet/index.html). Data analyzed in Section 5 are publicly available through dbGaP (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000222.v3.p2).

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors wish to thank the NIH dbGaP data repository (accession number phs000222.v3.p2). *Conflict of Interest*: None declared.

Funding

National Institutes of Health (NIH) (Grants NS073671, NS082062, NS036630, and GM124104).

References

Alexander-Bloch, A., Giedd, J. N. *and others*. (2013). Imaging structural co-variance between human brain regions. *Nature Reviews Neuroscience* **14**, 322–336.

Biglan, K. M., Zhang, Y., Long, J. D., Geschwind, M., Kang, G. A., Killoran, A., Lu, W., McCusker, E., Mills, J. A., Raymond, L. A. *and others*. (2013). Refining the diagnosis of Huntington disease: the PREDICT-HD study. *Frontiers in Aging Neuroscience* **5**:12. doi: 10.3389/fnagi.2013.00012.

Bullmore, E. T. and Bassett, D. S. (2011). Brain graphs: graphical models of the human brain connectome. *Annual Review of Clinical Psychology* **7**, 113–140.

Chen, Z. J., He, Y., Rosa-Neto, P., Germann, J. and Evans, A. C. (2008). Revealing modular architecture of human brain structural networks by using cortical thickness from MRI. *Cerebral Cortex* **18**, 2374–2381.

Cuingnet, R., Glaunès, J. A., Chupin, M., Benali, H. and Colliot, O. (2013). Spatial and anatomical regularization of SVM: a general framework for neuroimaging data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 682–696.

Desikan, R. S, Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T. *and others*. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980.

Eidelberg, D., Surmeier, D. J. *and others*. (2011). Brain networks in Huntington disease. *The Journal of Clinical Investigation* **121**, 484–492.

Engler, D. and Li, Y. (2009). Survival analysis with high-dimensional covariates: an application in microarray studies. *Statistical Applications in Genetics and Molecular Biology* **8**, 1–22.

Fan, J., Lin, H., Zhou, Y. *and others*. (2006). Local partial-likelihood estimation for lifetime data. *The Annals of Statistics* **34**, 290–325.

Feigin, A., Tang, C., Ma, Y., Mattis, P., Zgaljardic, D., Guttman, M., Paulsen, J. S., Dhawan, V. and Eidelberg, D. (2007). Thalamic metabolism and symptom onset in preclinical Huntington's disease. *Brain* **130**, 2858–2867.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546.

Hua, X., Ching, C. R. K., Mezher, A., Gutman, B. A., Hibar, D. P., Bhatt, P., Leow, A. D., Jack, C. R., Bernstein, M. A., Weiner, M. W. *and others*. (2016). MRI-based brain atrophy rates in ADNI phase 2: acceleration and enrichment considerations for clinical trials. *Neurobiology of Aging* **37**, 26–37.

Huang, J., Ma, S., Li, H. and Zhang, C.-H.. (2011). The sparse laplacian shrinkage estimator for high-dimensional regression. *Annals of Statistics* **39**, 2021.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27–30.

Kunegis, J., Schmidt, S., Lommatzsch, A., Lerner, J., De Luca, E. W. and Albayrak, S. (2010). Spectral analysis of signed graphs for clustering, prediction and visualization. In: Parthasarathy, S. *and others* (editors), *Proceedings of the 2010 SIAM International Conference on Data Mining*. Columbus, Ohio: SIAM, pp. 559–570.

Lee, J.-M., Wheeler, V. C., Chao, M. J., Vonsattel, J. P. G., Pinto, R. M., Lucente, D., Abu-Elneel, K., Ramos, E. M., Mysore, J. S., Gillis, T. *and others*. (2015). Identification of genetic factors that modify clinical onset of Huntington's disease. *Cell* **162**, 516–526.

Li, C. and Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The Annals of Applied Statistics* **4**, 1498.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Annals of Statistics* **36**, 261.

Liu, X. and Zeng, D. (2013). Variable selection in semiparametric transformation models for right-censored data. *Biometrika*, ast029.

MacDonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S. A., James, M., Groot, N. *and others*. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983.

Novak, M., Seunarine, K., Gibbard, C., Clark, C. and Tabrizi, S. J. (2012). G09 structural connectivity-based topography of the basal ganglia is altered in premanifest and early manifest Huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry* **83**, A29–A29.

Paulsen, J. S., Long, J. D., Johnson, H. J., Aylward, E. H., Ross, C. A., Williams, J. K., Nance, M. A., Erwin, C. J., Westervelt, H. J., Harrington, D. L. *and others*. (2014). Clinical and biomarker changes in premanifest Huntington disease show trial feasibility: a decade of the PREDICT-HD study. *Frontiers in Aging Neuroscience* **6**:78.

Rosas, H. D., Goodman, J., Chen, Y. I., Jenkins, B. G., Kennedy, D. N., Makris, N., Patti, M., Seidman, L. J., Beal, M. F. and Koroshetz, W. J. (2001). Striatal volume loss in HD as measured by MRI and the influence of CAG repeat. *Neurology* **57**, 1025–1028.

ROSS, C. A., AYLWARD, E. H., WILD, E. J., LANGBEHN, D. R., LONG, J. D., WARNER, J. H., SCAHILL, R. I., LEAVITT, B. R., STOUT, J. C., PAULSEN, J. S. *and others*. (2014). Huntington disease: natural history, biomarkers and prospects for therapeutics. *Nature Reviews Neurology* **10**, 204–216.

SUN, H., LIN, W., FENG, R. AND LI, H. (2014). Network-regularized high-dimensional cox regression for analysis of genomic data. *Statistica Sinica* **24**, 1433.

TZOURIO-MAZOYER, N., LANDEAU, B., PAPATHANASSIOU, D., CRIVELLO, F., ETARD, O., DELCROIX, N., MAZOYER, B. AND JOLIOT, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**, 273–289.

ZHANG, B. AND HORVATH, S.. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* **4**: Article17.

ZHANG, H. H. AND LU, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.

ZHANG, W., JOHNSON, N., WU, B. AND Kuang, R. (2012). Signed network propagation for detecting differential gene expressions and DNA copy number variations. In: *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. Orlando, Florida: ACM, pp. 337–344.

ZHANG, W., OTA, T., SHRIDHAR, V., CHIEN, J., WU, B. AND KUANG, R. (2013). Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Computational Biology* **9**, e1002975.

ZHANG, Y., LONG, J. D., MILLS, J. A., WARNER, J. H., LU, W. AND PAULSEN, J. S. (2011). Indexing disease progression at study entry with individuals at-risk for Huntington disease. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **156**, 751–763.

[*Received November 14, 2017; revised April 3, 2018; accepted for publication April 22, 2018*]